

TO WEIGHT OR NOT TO WEIGHT? BALANCING INFLUENCE OF INITIAL ITEMS IN ADAPTIVE TESTING

HUA-HUA CHANG

UNIVERSITY OF ILLINOIS

ZHILIANG YING

COLUMBIA UNIVERSITY

It has been widely reported that in computerized adaptive testing some examinees may get much lower scores than they would normally if an alternative paper-and-pencil version were given. The main purpose of this investigation is to quantitatively reveal the cause for the underestimation phenomenon. The logistic models, including the 1PL, 2PL, and 3PL models, are used to demonstrate our assertions. Our analytical derivation shows that, under the maximum information item selection strategy, if an examinee failed a few items at the beginning of the test, easy but more discriminating items are likely to be administered. Such items are ineffective to move the estimate close to the true θ , unless the test is sufficiently long or a variable-length test is used. Our results also indicate that a certain weighting mechanism is necessary to make the algorithm rely less on the items administered at the beginning of the test.

Key words: computerized adaptive testing, MLE, Fisher information, a -stratified method, item selection algorithm

1. Introduction

It has been widely speculated that in computerized adaptive testing (CAT) some examinees may get much lower scores than they would normally if an alternative paper-and-pencil (P&P) version were taken. As evidence of this, in 2000 Educational Testing Service (ETS) found that the GRE CAT system did not produce reliable scores for about a few thousand test takers. ETS offered them a chance to retake the test at no charge (Carlson, 2000). *Business Week* (Merritt, 2003) reported that in 2002, ETS incorrectly scored nearly a thousand students' CAT-GMATs, potentially affecting the chances of would-be MBAs getting into top-tier schools. Given that most CATs are high-stakes examinations, improving their reliability has become urgent.

To facilitate a remedy, it is necessary to identify potential "flaws" in the scoring formula in the CAT designs. The objective of this investigation is to psychometrically reveal what is most likely to account for the underestimation phenomenon. Since obtaining a real data set used during the incidents may not be possible, our effort has been limited to analytical assessment. The logistic models, including the 1PL, 2PL, and 3PL models, are used to demonstrate our assertions. Our analytical result shows that the maximum information method tends to select items with the highest a -parameters, which may cause a big step size in θ estimation at the beginning of the test. Consequently, it is plausible that if an examinee misses a number of initial items and the test length is short to moderate, then he or she may not be able to regain a score close to the true θ . On the other hand, it is also possible that a person who guesses correctly early in the test could

This research was partially supported by the NSF Grants SES0241020 and SES0613025. The authors thank the Editor, Associate Editor and two anonymous reviewers for their comments and suggestions. Send further information to Hua-Hua Chang, Department of Psychology, 603 E. Daniel Street, M/C 716, Champaign, IL 61820.

Requests for reprints should be sent to Hua-Hua Chang, University of Illinois, Urbana, USA. E-mail: hhchang@uiuc.edu

be overestimated. Our results imply that a certain weighting mechanism is necessary to make the algorithm rely less on the items administered at the beginning of the test. For this purpose, the a -stratified item selection method proposed by Chang and Ying (1999) demonstrated a significant improvement in estimation stability. This paper is organized as the following: The main analytic results are presented in the next section. Section 3 summarizes findings from numerical studies. The last section gives some additional remarks.

2. Main Results

Let Y_i be the score for a randomly selected examinee on the i th item; $Y_i = 1$ if the answer is correct and $Y_i = 0$ if incorrect. Let $Y_i = 1$ with probability $P_i(\theta)$ and $Y_i = 0$ with probability $1 - P_i(\theta)$, where $P_i(\theta)$ denotes the probability of a correct response for a randomly chosen examinee of latent trait θ , where θ is unknown and has the domain $(-\infty, \infty)$ or some subinterval on $(-\infty, \infty)$. When the three-parameter logistic model (3PL) is used, the probability becomes

$$P_i(\theta) = c_i + (1 - c_i) \frac{1}{1 + e^{-a_i(\theta - b_i)}}, \quad (1)$$

where a_j is the item discrimination parameter, b_j is the difficulty parameter, and c_j is the guessing parameter. There are two special cases of (1): the 1PL model (when $c_i \equiv 0$ and $a_i \equiv 1$) and the 2PL model (when $c_i \equiv 0$).

Suppose that an examinee with a fixed θ is given n items Y_1, Y_2, \dots, Y_n . Then θ can be estimated by maximizing the likelihood function

$$L_n(\theta) = \prod_{i=1}^n P_i(\theta)^{Y_i} Q_i(\theta)^{1-Y_i}, \quad (2)$$

where $P_i(\theta)$ is item response function and $Q_i(\theta) = 1 - P_i(\theta)$. Let $\hat{\theta}_n$ denote the resulting estimator: $\hat{\theta}_n$ solves the following maximum likelihood estimating equation

$$U_n(\theta) = \frac{\partial}{\partial \theta} \log L_n(\theta) = \sum_{i=1}^n \left\{ \frac{\partial}{\partial \theta} \log \frac{P_i(\theta)}{Q_i(\theta)} \right\} [Y_i - P_i(\theta)] = 0. \quad (3)$$

Note that in (3) $Y_i - P_i(\theta)$ is “observed”–“expected”, and hence it has mean 0. Thus $U_n(\theta)$ is a weighted sum of “observed”–“expected” with weights

$$w_i = \frac{\partial}{\partial \theta} \left\{ \log \frac{P_i(\theta)}{Q_i(\theta)} \right\}, \quad i = 1, \dots, n.$$

It is well known that, under suitable regularity conditions, $\hat{\theta}_n$ is asymptotically normal, centered at the true θ with variance approximated by $I_{(n)}^{-1}(\hat{\theta}_n)$, where $I_{(n)}(\theta)$ is the Fisher test information function. One original motivation for CAT is to maximize the Fisher information so that $\hat{\theta}_n$ will be most accurate. This can be achieved by recursively estimating θ with currently available data and assigning further items adaptively.

In order to promote remedies for the underestimation problem, the sensitivity of $\hat{\theta}_n$ for small n needs to be investigated. Let us first consider the 1PL model. Without loss of generality, we assume the common a parameter to be 1. Since $\log \frac{P_i(\theta)}{Q_i(\theta)} = \theta - b_i$, therefore $\frac{\partial}{\partial \theta} \left\{ \log \frac{P_i(\theta)}{Q_i(\theta)} \right\} = 1$.

According to (3), the likelihood estimation function takes the form in (4) after n items were administered:

$$U_n(\theta) = \sum_{i=1}^n \left(Y_i - \frac{e^{\theta-b_i}}{1+e^{\theta-b_i}} \right). \quad (4)$$

Note that for the MLE $\hat{\theta}_n$, $U_n(\hat{\theta}_n) = 0$. Let b_{n+1} be the item difficulty parameter for the $(n+1)$ th item selected. Let $\hat{\theta}_{n+1}$ be the maximum likelihood estimator (MLE) based on $n+1$ items, i.e., $\hat{\theta}_{n+1}$ solves $U_{n+1}(\hat{\theta}_{n+1}) = 0$. By the mean-value theorem,

$$U_{n+1}(\hat{\theta}_{n+1}) - U_{n+1}(\hat{\theta}_n) = \frac{\partial U_{n+1}(\theta_{n+1}^*)}{\partial \theta} (\hat{\theta}_{n+1} - \hat{\theta}_n), \quad (5)$$

where θ_{n+1}^* lies between $\hat{\theta}_{n+1}$ and $\hat{\theta}_n$. Note that (5) also holds for the 2PL and 3PL models. For the 1PL model, it can be verified that

$$U_{n+1}(\hat{\theta}_{n+1}) - U_{n+1}(\hat{\theta}_n) = -I_{(n+1)}(\theta_{n+1}^*) (\hat{\theta}_{n+1} - \hat{\theta}_n), \quad (6)$$

where $I_{(n+1)}(\theta) = \sum_{i=1}^{n+1} e^{\theta-b_i} / (1+e^{\theta-b_i})^2$ is the Fisher test information and is also equal to $-\frac{\partial U_{n+1}(\theta)}{\partial \theta}$. Since $U_n(\hat{\theta}_n) = 0$ for all fixed n , it can be shown

$$\hat{\theta}_{n+1} = \hat{\theta}_n + \frac{U_{n+1}(\hat{\theta}_n)}{I_{(n+1)}(\theta_{n+1}^*)}. \quad (7)$$

Since $U_{n+1}(\hat{\theta}_n) = U_n(\hat{\theta}_n) + (Y_{n+1} - \frac{e^{\hat{\theta}_n-b_{n+1}}}{1+e^{\hat{\theta}_n-b_{n+1}}})$, therefore the following recursion holds:

$$\hat{\theta}_{n+1} = \hat{\theta}_n + \frac{1}{I_{(n+1)}(\theta_{n+1}^*)} \left(Y_{n+1} - \frac{e^{\hat{\theta}_n-b_{n+1}}}{1+e^{\hat{\theta}_n-b_{n+1}}} \right). \quad (8)$$

For the 1PL model, the relationship between $\hat{\theta}_n$ and $\hat{\theta}_{n+1}$ described in (8) indicates that, if the item pool is sufficiently rich that allows each θ to match closely to a difficulty parameter b , then $b_{n+1} \approx \hat{\theta}_n$ and $e^{\hat{\theta}_n-b_{n+1}} / (1+e^{\hat{\theta}_n-b_{n+1}}) \approx \frac{1}{2}$. This implies that the one-step update from $\hat{\theta}_n$ to $\hat{\theta}_{n+1}$ is $\pm \frac{1}{2}$ divided by $I_{(n+1)}(\theta_{n+1}^*)$, which is typical of order $2/n$ for large n . Consequently, the larger the n is, the smaller the one-step adjustment it gets. It is conceivable that if the examinee misses a couple of items at the beginning of the test, the one-step update would push the examinee toward the negative direction too quickly for the 1PL model.

Now let's consider the 2PL model. It can be verified that $\frac{\partial}{\partial \theta} \{\log \frac{P_i(\theta)}{Q_i(\theta)}\} = a_i$. According to (3), the estimation function becomes

$$U_n(\theta) = \sum_{i=1}^n a_i \left(Y_i - \frac{e^{a_i(\theta-b_i)}}{1+e^{a_i(\theta-b_i)}} \right), \quad (9)$$

and the test information function becomes

$$I_{(n)}(\theta) = \sum_{i=1}^n a_i^2 \frac{e^{a_i(\theta-b_i)}}{[1+e^{a_i(\theta-b_i)}]^2}. \quad (10)$$

According to (5), (9), and (10), by using the same reason to get (8), the following recursion holds:

$$\hat{\theta}_{n+1} = \hat{\theta}_n + \frac{a_{n+1}}{I_{(n+1)}(\theta_{n+1}^*)} \left(Y_{n+1} - \frac{e^{a_{n+1}(\hat{\theta}_n-b_{n+1})}}{1+e^{a_{n+1}(\hat{\theta}_n-b_{n+1})}} \right), \quad (11)$$

where θ_{n+1}^* is a point between $\hat{\theta}_{n+1}$ and $\hat{\theta}_n$. The situation becomes even more interesting for the 2PL model. Equation (11) indicates that the one-step update from $\hat{\theta}_n$ to $\hat{\theta}_{n+1}$ is $\pm 1/2$ multiplied by $a_{n+1} I_{(n+1)}^{-1}(\theta_{n+1}^*)$, which indicates that the size of the step may be determined by the value of a for small n . Consequently, the larger the n is, the smaller the one-step adjustment it gets. As indicated by many authors, the maximum information approach would select the items with the highest a -values, which may cause a big step size at the beginning of the test. Therefore, it is plausible that if the examinee misses a number of initial items and the test length is short to moderate, then he or she may not be able to regain a score (estimate) comparable (close) to the true θ , even though he or she responds well to the rest of the items.

When the 3PL model is used, the recursion can be approximated:

$$\hat{\theta}_{n+1} \approx \hat{\theta}_n + \frac{a_{n+1}}{I_{(n+1)}(\theta_{n+1}^*)} g(c_{n+1})(Y_{n+1} - P_{n+1}(\hat{\theta}_n)), \quad (12)$$

where $2/3 \leq g(c_{n+1}) \leq 1$, θ_{n+1}^* lies between $\hat{\theta}_{n+1}$ and $\hat{\theta}_n$, $I_{(n+1)}(\theta)$ is the Fisher test information function, and $P_{n+1}(\theta)$ is ICC for the $(n+1)$ th item. (See Appendix for a theorem.) Since $g(c_{n+1})$ is between $2/3$ and 1 , the discussion about the relationship between the value of a -parameter and the step size should be the same as that for the 2PL model.

Chang and Ying (1996, 1999) argued that the a -parameter should be selected in an ascending order. Their motivations come from the considerations of efficiency improvement and item exposure balance. In view of (11) and (12), an additional benefit of the a -stratified approach is that it automatically adjusts step sizes in updating the current estimation of θ . In particular, it shrinks weights at early stages, making it less likely to have extreme values in estimating θ . It also inflates weights at final stages, counteracting the effect of the multiplier $I_{(n+1)}^{-1}(\theta_{n+1}^*)$ and making it more likely to adjust the final estimator of θ . It is clear that the ascending order of a_n as advocated by Chang and Ying plays a pivotal role, ensuring higher efficiency, giving more balanced exposure rates, reducing fluctuation due to initial item response irregularity, and increasing effectiveness in counteracting initial item influence by responses to later items.

Therefore, it is plausible that if the examinee misses a number of initial items and the test length is short, then he or she may not be able to regain a score (estimate) comparable (close) to the true θ , even though he or she responds well to the rest of the items. According to (11) and (12), it is also possible that a person who guesses correctly early in the test could be overestimated. Actually, "Pay extra attention to first few questions" was advised by several GRE preparation books (e.g., Kaplan, 2004). In order to overcome the problem, the item selection strategy needs to be adjusted so that it selects items with low discrimination parameters at the beginning of the test.

3. Simulation Studies

A pilot simulation study was conducted to get numerical evidence supporting our theoretical findings.

Item Pool Structure Assume the item pool is so sufficiently rich that for every given θ value one can find a corresponding difficulty parameter b with the same value. The item pool was partitioned into four strata and the discrimination parameter was identical within each stratum, with a equaled to 0.5, 1.0, 1.5, and 2.0, respectively, for the four strata. Without jeopardizing the generalizability of the findings, all discrimination parameters within each stratum were kept constant to maintain a clear change in item parameter distribution as testing progressed. Within each stratum, the values of the difficulty b -parameters can be generated with the same value of θ . For simplicity, the guessing c parameters for all items were set at zero.

Latent Trait Distribution One thousand examinees were included with a fixed θ value at 0. Note that the true θ can be fixed at any point.

Test length and termination rule The test length examined was 40 items. Though it is a bit longer than most adaptive tests, we deliberately chose 40 to show the trend more clearly.

Item Selection Rules Hau and Chang (2001) made a comparison of efficiency in terms of MSE among several methods and reported that the ascending a -method was better than the descending a -method. The focus of this research is to examine whether the use of more discriminating items at the beginning of testing would cause “convergence to a wrong point.” Two approaches, the descending a - and ascending a -methods, were used in the simulation study. In the descending a -method, the use of large a -parameter items were in the reverse order of the ascending a -method, that is, large a -parameter items were used first followed by small a ones. The objective of the research is to defend a general principle—low- a items should be used first and high- a items should be used last, thus it will not entail specific item-selection methods, such as the maximum information method, the a -stratified method, etc. In this regard the simulation design is essentially for all MLE based procedures that can be portrayed by (11).

- Step 1. The item pool is partitioned into four strata by the a -parameter, with the first and last strata containing, respectively, the smallest and the largest- a items.
- Step 2. Accordingly, the testing process is also partitioned into four stages to match the four item strata.
- Step 3. At the k th stage, 10 items are selected from the k th stratum. The test-taker’s ability is updated by (11), which is equivalent to maximizing the likelihood function constructed from the responses to the items already taken. Then items of difficulty parameter equal to the estimated ability are selected and administered as the next item.
- Step 4. Step 3 is repeated for $k = 1$ through $k = 4$ stages.

The steps in the descending a -method are:

- Step 1. The item pool is partitioned into four strata by the a -parameter as in the ascending a -method. However, contrary to the ascending a -method, the earlier and latter strata now contain, respectively, the higher and lower a items.
- Steps 2 to 4. Identical to the steps in the ascending a -method, the entire testing process is also partitioned into four stages to match the four item strata.

Initial Estimators Let $\hat{\theta}_1$ be the initial estimator of θ . Eleven initial estimation points were selected: $\hat{\theta}_1 = -3.0, -2.5, -2.0, -1.5, -1.0, 0, 1.0, 1.5, 2.0, 2.5,$ and 3.0 . Note: the true $\theta = 0.0$. The reason of using different initial estimators is to assess the variability of estimation accuracy possibly caused by different initial performance among examinees.

Evaluation Criterion Average bias (Bias), mean squared error (MSE), and average number of correct (ANC) were used as evaluation criterion, which are defined in the following:

$$\text{Bias} = \frac{1}{1000} \sum_{j=1}^{1000} [\hat{\theta}(j) - \theta_0], \quad (13)$$

$$\text{MSE} = \frac{1}{1000} \sum_{j=1}^{1000} [\hat{\theta}(j) - \theta_0]^2, \quad (14)$$

and

$$\text{ANC} = \frac{1}{1000} \sum_{j=1}^{1000} \sum_{i=1}^{40} Y_{ij}, \quad (15)$$

where $\hat{\theta}(j)$ is the final estimator for the j th examinee, θ_0 is the true ability which is set to 0 in our simulation, and Y_{ij} is the observed score on the i th item for examinee j .

Results Figure 1 shows the mean squared errors for the two methods. For each method, MSEs were calculated based on 1000 replications at each of the 11 starting points. The graph represents the MSEs as a function of the starting values. Note that $\theta = 0$ represents the true latent trait value in our simulation. For the descending a -method, the larger the difference between the initial estimator and the true θ is, the higher is the value of MSE. The same pattern was found in the bias plots in Figure 2. If the initial value is lower than the true θ , the final estimator is negatively biased. On the other hand, the final estimator will be positively biased if the initial estimator is higher than the true θ . These figures clearly indicate that the ascending a -method generated much more consistent results for both bias and MSE. The simulation results may also imply that if the item selection algorithm relies on items with the highest a -parameter values at the beginning of the test, it is plausible that if the examinee misses a number of initial items, then he/she may not be able to regain an estimated $\hat{\theta}$ that is comparable to the true θ , even though he/she responds well to the rest of the items. For instance, according to Figure 2, the average bias of the descending method at -3 was -1.8 , however, according to Figure 3, the corresponding average number of correct items was 37. The simulation results clearly indicate overestimation is plausible. According to (11), it is possible that a person who guesses correctly early in the test could be overestimated.

4. Conclusions

In this paper both the analytical derivations and the empirical simulation study showed that under the maximum information item-selection strategy, if an examinee failed a few items at the beginning of a test, easy (but more discriminating) items are likely to be administered. And such items are ineffective to bring the estimate close to the true θ , unless either the test length is sufficiently long or a variable-length test is used. In the 2000 GRE incident, even though ETS refused to comment on whether the examinees who were offered to retake the GRE were scored lower or higher (Carlson, 2000), our speculation is that they most likely received extremely low scores. According to (11) and (12), it is likely that in 2000 about the same number of examinees were scored higher than what they deserved.

The derived recursions, in particular under the mathematically straightforward 2PL model, indicate that the estimation updates based on high-discriminating items may lead the CAT off target at the beginning of a test. The small-scale simulation study provided further support to our assertion. The results presented in this paper may help us to design more robust item-selection algorithms. In view of (11), the a -stratified approach is a promising alternative in that it automatically adjusts step sizes when updating the current estimation of θ . It shrinks the weights at early stages, making it less likely to have extreme values in estimating θ . It also inflates the weights at final stages, counteracting the effect of the multiplier $I_{(n+1)}^{-1}(\theta_{n+1}^*)$, and makes it more efficient to adjust the final estimator of θ . The a -stratified method is only one solution, and other promising solutions might come along, including incorporating the maximum information method with certain weighting mechanisms. Undeniably, adopting an item-selection strategy that places less weight on the items administered at the beginning of the testing will greatly increase

MSE

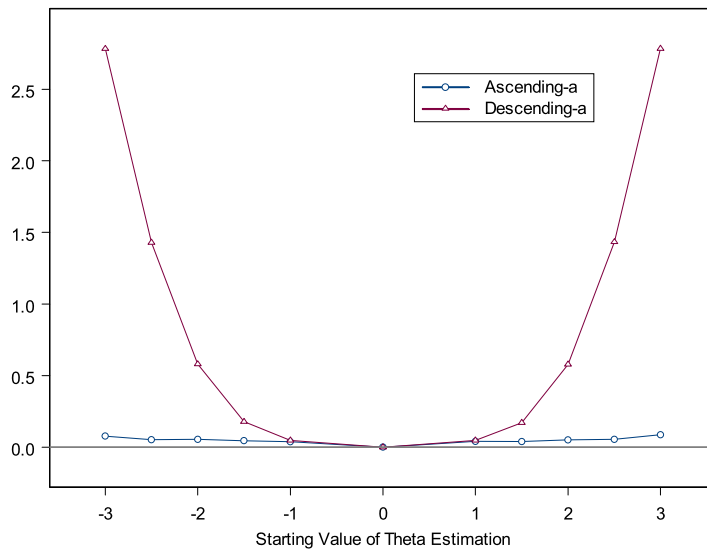


FIGURE 1.
MSEs of the two methods.

Bias

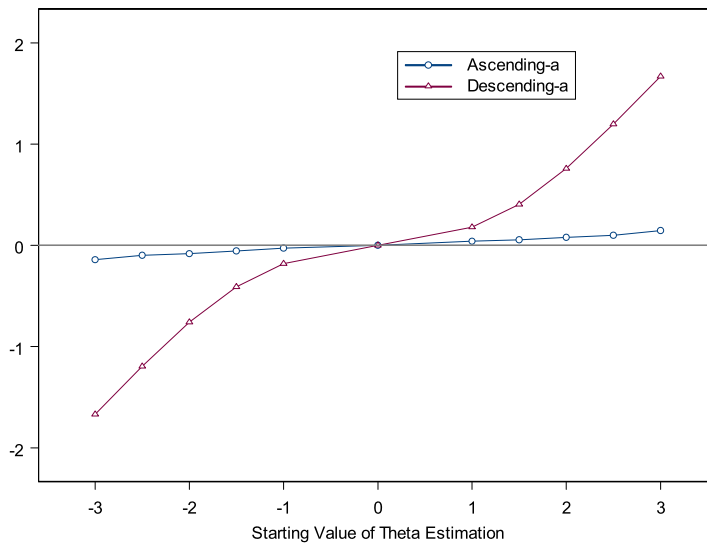


FIGURE 2.
Biases for the two methods.

the reliability of a CAT. Finally, the results presented in this paper may imply that, despite some shortcomings, CAT undoubtedly has a great future because new developments in psychometric theory will enable us to solve the problems encountered in current large-scale applications.

Average Number Correct

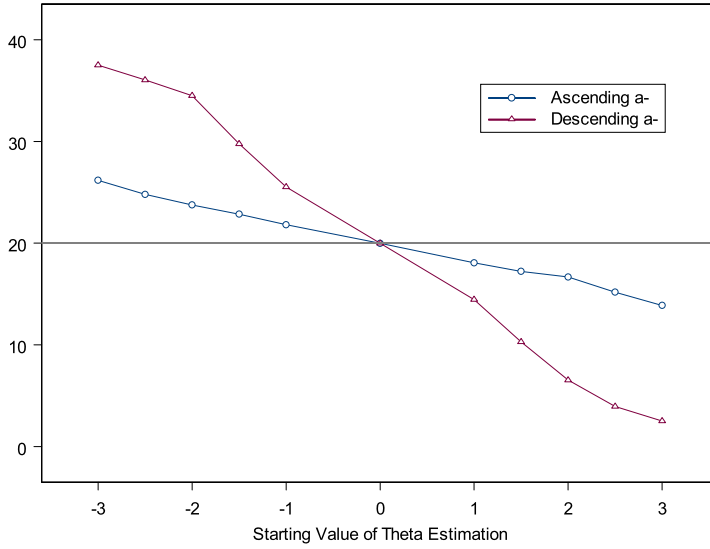


FIGURE 3.
Average number of correct items of the two methods.

Appendix

Theorem. Assume the 3PL model and the maximum information item-selection method is used. Let a_{n+1} , b_{n+1} , and c_{n+1} are the item parameters for the $(n + 1)$ th item which are sequentially selected. Let $\hat{\theta}_{n+1}$ and $\hat{\theta}_n$ MLEs based on $n + 1$ items and n items, respectively. The relationship between $\hat{\theta}_{n+1}$ and $\hat{\theta}_n$ can be expressed by

$$\hat{\theta}_{n+1} \approx \hat{\theta}_n + \frac{a_{n+1}}{I_{(n+1)}(\theta_{n+1}^*)} g(c_{n+1}) \left(Y_{n+1} - \left(c_{n+1} + (1 - c_{n+1}) \frac{e^{a_{n+1}(\hat{\theta}_n - b_{n+1})}}{1 + e^{a_{n+1}(\hat{\theta}_n - b_{n+1})}} \right) \right), \quad (16)$$

where $2/3 \leq g(c_{n+1}) \leq 1$, θ_{n+1}^* lies between $\hat{\theta}_{n+1}$ and $\hat{\theta}_n$, and $I_{(n+1)}(\theta)$ is the Fisher test information function.

Proof: By the mean-value theorem,

$$U_{n+1}(\hat{\theta}_{n+1}) - U_{n+1}(\hat{\theta}_n) = \frac{\partial}{\partial \theta} U_{n+1}(\theta_{n+1}^*) [\hat{\theta}_{n+1} - \hat{\theta}_n], \quad (17)$$

where θ_{n+1}^* is between $\hat{\theta}_{n+1}$ and $\hat{\theta}_n$. For the 3PL model it can be verified that

$$\frac{\partial}{\partial \theta} \left\{ \log \frac{P_i(\theta)}{Q_i(\theta)} \right\} = \frac{a_i e^{a_i(\theta - b_i)}}{c_i + e^{a_i(\theta - b_i)}}.$$

According to (3) the likelihood equation is

$$U_n(\theta) = \sum_{i=1}^n \frac{a_i e^{a_i(\theta - b_i)}}{c_i + e^{a_i(\theta - b_i)}} \left(Y_i - c_i - (1 - c_i) \frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}} \right) = 0, \quad (18)$$

where c_i is the guessing parameter for item i . Since $\hat{\theta}_{n+1}$ and $\hat{\theta}_n$ are MLE, $U_{n+1}(\hat{\theta}_{n+1}) = 0$ and $U_n(\hat{\theta}_n) = 0$. Since

$$U_{n+1}(\hat{\theta}_n) = U_n(\hat{\theta}_n) + \frac{a_{n+1}e^{a_{n+1}(\hat{\theta}_n - b_{n+1})}}{c_{n+1} + e^{a_{n+1}(\hat{\theta}_n - b_{n+1})}} \left(Y_{n+1} - c_{n+1} - (1 - c_{n+1}) \frac{e^{a_{n+1}(\hat{\theta}_n - b_{n+1})}}{1 + e^{a_{n+1}(\hat{\theta}_n - b_{n+1})}} \right),$$

(17) becomes

$$\begin{aligned} & - \frac{a_{n+1}e^{a_{n+1}(\hat{\theta}_n - b_{n+1})}}{c_{n+1} + e^{a_{n+1}(\hat{\theta}_n - b_{n+1})}} \left(Y_{n+1} - c_{n+1} - (1 - c_{n+1}) \frac{e^{a_{n+1}(\hat{\theta}_n - b_{n+1})}}{1 + e^{a_{n+1}(\hat{\theta}_n - b_{n+1})}} \right) \\ & = \frac{\partial}{\partial \theta} U_{n+1}(\theta_{n+1}^*) [\hat{\theta}_{n+1} - \hat{\theta}_n]. \end{aligned} \quad (19)$$

Solve $\hat{\theta}_{n+1}$ from (19)

$$\hat{\theta}_{n+1} = \hat{\theta}_n - \frac{1}{\frac{\partial}{\partial \theta} U_{n+1}(\theta_{n+1}^*)} \times \frac{a_{n+1}e^{a_{n+1}(\hat{\theta}_n - b_{n+1})}}{c_{n+1} + e^{a_{n+1}(\hat{\theta}_n - b_{n+1})}} (Y_{n+1} - P_{n+1}(\hat{\theta}_n)). \quad (20)$$

Now we need to compute $\frac{\partial}{\partial \theta} U_{n+1}(\theta_{n+1}^*)$. Note that the item parameters in (18) are sequentially selected so that the Fisher item information of each item is maximized. According to Lord (1980, p. 152), the Fisher item information reaches its maximum when

$$b \cong \theta - \frac{1}{a} \log \frac{1 + \sqrt{1 + 8c}}{2}. \quad (21)$$

Therefore, when the choice of the difficulty parameter satisfies (21), it is easy to see that the weights in (18)

$$\frac{a_i e^{a_i(\theta - b_i)}}{c_i + e^{a_i(\theta - b_i)}} \cong \frac{a_i(1 + \sqrt{1 + 8c_i})}{2c_i + 1 + \sqrt{1 + 8c_i}}.$$

Therefore, an approximation to $U_n(\theta)$ is

$$\sum_{i=1}^n \frac{a_i(1 + \sqrt{1 + 8c_i})}{2c_i + 1 + \sqrt{1 + 8c_i}} \left(Y_i - c_i - (1 - c_i) \frac{e^{a_i(\theta - b_k)}}{1 + e^{a_i(\theta - b_k)}} \right). \quad (22)$$

By taking derivative on $Y_i - P_i(\theta)$, we have

$$\frac{\partial}{\partial \theta} U_{n+1}(\theta_{n+1}^*) \approx - \sum_{i=1}^{n+1} I_i(\theta_{n+1}^*). \quad (23)$$

Assume $(a_{n+1}, b_{n+1}, c_{n+1})$ and $\hat{\theta}_n$ satisfy (21), according to (20) and (23),

$$\hat{\theta}_{n+1} \approx \hat{\theta}_n + \frac{a_{n+1}}{I_{(n+1)}(\theta_{n+1}^*)} g(c_{n+1}) (Y_{n+1} - P_{n+1}(\hat{\theta}_n)), \quad (24)$$

where $g(c_{n+1}) = \frac{1 + \sqrt{1 + 8c_{n+1}}}{2c_{n+1} + 1 + \sqrt{1 + 8c_{n+1}}}$. Since $0 \leq c \leq 1$, it is obvious that $2/3 \leq g(c_{n+1}) \leq 1$. \square

PSYCHOMETRIKA

References

- Carlson, S. (2000). ETS finds flaws in the way online GRE rates some students. *Chronicle of Higher Education*, 47(8), A47.
- Chang, H.H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, 20, 213–229.
- Chang, H.H., & Ying, Z. (1999). A-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, 23(3), 211–222.
- Hau, K.-T., & Chang, H. (2001). Item selection in computerized adaptive testing: should more discriminating items be used first? *Journal of Educational Measurement*, 28, 249–266.
- Kaplan (2004). *GRE Exam* (2005 ed., p. 198). New York: Kaplan Publishing.
- Lord, M.F. (1980). *Applications of item response theory to practical testing problems*. Hillsdale: Laurence Erlbaum Associates, Publishers.
- Merritt, J. (2003). Why the folks at ETS flunked the course—a tech-savvy service will soon be giving B-school applicants their GMATs. *Business Week*, December 29, 2003.

Manuscript received 25 JAN 2007

Final version received 16 SEP 2007